Introduction aux Modèles Mixtes

UE Statistique de la science des données

Pierre Pudlo

1. Modèles Mixtes : Une Introduction

Introduction

Les modèles mixtes, aussi appelés modèles linéaires à effets mixtes (c'est-à-dire aléatoires et fixes), sont une généralisation des modèles linéaires classiques. Dans un modèle linéaire standard, on suppose que tous les effets sont fixes, c'est-à-dire qu'ils sont constants et applicables à l'ensemble de la population. Les modèles mixtes étendent cette approche en incluant des effets aléatoires, permettant ainsi de mieux représenter la variabilité entre différents groupes ou niveaux hiérarchiques. Cette flexibilité est particulièrement utile dans le cas de données hiérarchiques, corrélées, ou mesurées de manière répétée.

Modèle Linéaire Classique

Un modèle linéaire classique peut être écrit sous la forme suivante :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Où :

- $\mathbf{Y}$: Vecteur des réponses observées.
- $\mathbf{X}$: Matrice des prédicteurs (ou variables explicatives) pour les effets fixes.
- $\boldsymbol{\beta}$: Vecteur des paramètres des effets fixes.
- $\boldsymbol{\varepsilon}$: Erreur aléatoire, supposée suivre une distribution normale $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2\mathbf{I}_n)$.

Ce modèle est approprié lorsque l'on suppose que tous les effets sont identiques pour chaque observation, sans variabilité entre les groupes.

Modèle Mixte : Généralisation

Les modèles mixtes généralisent les modèles linéaires en ajoutant des **effets aléatoires** pour modéliser la variabilité entre les groupes. Un modèle mixte peut être représenté comme suit :

$$Y_{ij} = X_{ij}\beta + Z_{ij}u_j + \varepsilon_{ij}$$

Où :

- Y_{ij} : Réponse de l'observation i dans le groupe j.
- X_{ij} : Matrice ligne des prédicteurs de l'observation i dans le groupe j pour les effets fixes
- β : Matrice colonne des effets fixes (communs à tous les groupes).
- Z_{ij} : Matrice ligne des prédicteurs de l'observation i dans le groupe j pour les effets aléatoires.
- u_j : Matrice colonne des effets aléatoires associés au groupe j, supposé suivre une distribution normale $u_j \sim \mathcal{N}_q(0, G)$.
- ε_{ij} : Erreur aléatoire, supposée suivre une distribution normale $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

$$Y_{ij} = X_{ij}\beta + Z_{ij}u_j + \varepsilon_{ij}$$

Ainsi, la variance totale de $[Y_{ij}|X_{ij}, Z_{ij}]$ est composée de la variance des effets aléatoires (u_j) et de la variance des erreurs résiduelles (ε_{ij}). Cela permet de capturer la variabilité non expliquée par les effets fixes, mais liée aux différences entre les groupes.

Exemple Concret

Considérons une étude sur les performances scolaires des élèves dans différentes écoles. On souhaite modéliser le score des élèves en fonction de plusieurs facteurs tels que le nombre d'heures d'étude (effet fixe). Cependant, il est raisonnable de penser que les performances peuvent également varier d'une école à l'autre en raison de facteurs non observés (comme la qualité de l'enseignement ou des ressources).

Le modèle mixte dans ce cas serait :

$$\text{Score}_{ij} = \beta_0 + \beta_1 \times \text{HeuresÉtude}_{ij} + u_j + \varepsilon_{ij}$$

Où :

- β_0 : Intercept global (moyenne générale des scores).
- β_1 : Effet fixe d'une heure d'étude sur le score.
- u_j : Effet aléatoire pour l'école j, supposé suivre $u_j \sim \mathcal{N}(0, \sigma_u^2)$.
- ε_{ij} : Erreur aléatoire pour chaque élève, supposée suivre $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

Ici donc, $Z_{ij} = 1$ est une matrice ligne de taille 1×1.

$$\text{Score}_{ij} = \beta_0 + \beta_1 \times \text{HeuresÉtude}_{ij} + u_j + \varepsilon_{ij}$$

Dans cet exemple, les effets fixes (β_1) capturent l'influence du nombre d'heures d'étude sur le score, tandis que les effets aléatoires (u_j) capturent la variabilité spécifique à chaque école. Cela permet de mieux modéliser la structure des données et de tenir compte des différences entre les écoles.

2. Un premier exemple : Étude Longitudinale de la Pression Artérielle

Contexte de l'Étude

On souhaite modéliser la pression artérielle de patients en fonction de plusieurs facteurs explicatifs, tels que l'âge, le sexe, l'IMC (indice de masse corporelle), ainsi que l'effet du temps sur chaque patient. La pression artérielle de chaque patient est mesurée plusieurs fois au fil des ans, ce qui induit une corrélation entre les mesures répétées sur le même individu. De plus, il peut y avoir des effets globaux liés à des périodes spécifiques, comme des variations saisonnières.

Formulation Mathématique du Modèle

Un modèle mixte adapté à cette étude peut être représenté par l'équation suivante :

$$\text{Pression}_{it} = \beta_0 + \beta_1 \times \text{Âge}_{it} + \beta_2 \times \text{IMC}_{it} + \beta_3 \times \text{Sexe}_i + u_i + v_t + \varepsilon_{it}$$

Où :

- β_0 : Intercept global

- $\beta_1, \beta_2, \beta_3$: Effets fixes des variables explicatives (âge, IMC, sexe).
- u_i : Effet aléatoire associé au patient i, supposé suivre $u_i \sim \mathcal{N}(0, \sigma_u^2)$. Cet effet capture la variabilité individuelle des patients.
- v_t : Effet aléatoire associé au temps t, qui pourrait capturer des variations temporelles communes à tous les patients, telles que des effets saisonniers. Cet effet est supposé suivre $v_t \sim \mathcal{N}(0, \sigma_v^2)$.
- ε_{it} : Erreur aléatoire spécifique à chaque observation, supposée suivre $\varepsilon_{it} \sim \mathcal{N}(0, \sigma^2)$.

$$\text{Pression}_{it} = \beta_0 + \beta_1 \times \hat{\text{Age}}_{it} + \beta_2 \times \text{IMC}_{it} + \beta_3 \times \text{Sexe}_i + u_i + v_t + \varepsilon_{it}$$

Ce modèle est un exemple typique de **modèle mixte à effets aléatoires croisés**, avec des effets aléatoires pour les patients (u_i) et pour les périodes de mesure (v_t).

Interprétation des Effets

- **Effets Fixes (β)** : Les effets fixes quantifient l'influence des variables explicatives sur la pression artérielle. Par exemple, β_1 représente l'effet de l'âge sur la pression artérielle moyenne. Si β_1 est positif, cela signifie que la pression artérielle tend à augmenter avec l'âge, tout autre facteur (IMC, Sexe) étant fixé.
- **Effet Aléatoire Patient (u_i)** : L'effet u_i représente la variabilité spécifique à chaque patient qui n'est pas expliquée par les effets fixes. Cela permet de modéliser les différences individuelles entre les patients, comme des variations dues à des facteurs génétiques ou des habitudes de vie non mesurées.
- **Effet Aléatoire Temps (v_t)** : L'effet v_t représente les variations dans la pression artérielle qui sont partagées par tous les patients à un moment donné, comme les variations saisonnières ou les effets de certaines politiques de santé.

Avantages de ce Modèle Mixte plus Complexe

- **Prise en compte de la Variabilité Intra- et Inter-individuelle** : Les modèles mixtes permettent de capturer à la fois les différences entre les patients et les variations propres à chaque patient dans le temps.
- **Flexibilité** : L'inclusion d'effets aléatoires à effets aléatoires croisés rend le modèle très flexible pour les données hiérarchiques ou longitudinales.

3. Un second exemple, sur le taux de cholestérol et l'activité physique

Introduction

Les modèles mixtes permettent non seulement de modéliser la variabilité inter- et intra-groupe, mais aussi d'étudier l'effet d'une covariable de manière plus nuancée. Un effet mixte sur une covariable implique que l'influence de cette covariable peut varier aléatoirement d'un groupe à l'autre. Cela est particulièrement utile pour modéliser des relations qui ne sont pas uniformes à travers différents groupes ou individus.

Nous allons explorer un exemple concret où l'effet d'une covariable est modélisé comme un effet mixte, offrant une compréhension plus complète de la variabilité au sein des données.

Contexte de l'Étude

Considérons une étude sur l'impact de l'activité physique sur le taux de cholestérol chez les patients appartenant à différents centres médicaux. L'effet de l'activité physique peut varier d'un centre à l'autre en raison de facteurs tels que l'encadrement, les infrastructures disponibles, ou les caractéristiques locales des patients.

Nous souhaitons modéliser le taux de cholestérol (Y_{ij}) en fonction des heures hebdomadaires d'activité physique (X_{ij}), mais nous pensons que l'activité physique pourrait ne pas être la même dans chaque centre médical, ce qui implique des effets différents du temps d'activité physique par centre.

Formulation Mathématique du Modèle

Le modèle mixte avec effet sur la covariable peut être écrit comme suit :

$$Y_{ij} = \beta_0 + (\beta_1 + u_j)X_{ij} + \epsilon_{ij}$$

Où :

- Y_{ij} : Taux de cholestérol du patient i dans le centre j.
- β_0 : Intercept global
- β_1 : Effet fixe global de l'activité physique sur le taux de cholestérol.
- u_j : Effet aléatoire associé au centre j, supposé suivre $u_j \sim \mathcal{N}(0, \sigma_u^2)$. Cet effet modélise la variabilité de l'activité physique entre les centres, donc de leur effet.

- X_{ij} : Nombre d'heures d'activité physique du patient i dans le centre j.
- ϵ_{ij} : Erreur aléatoire pour chaque observation, supposée suivre $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

Dans ce modèle, l'effet de la covariable (X_{ij}) est modifié par l'effet aléatoire u_j. Cela signifie que la pente de la relation entre l'activité physique et le taux de cholestérol peut varier d'un centre à l'autre.

Interprétation des Effets

$$Y_{ij} = \beta_0 + (\beta_1 + u_j)X_{ij} + \epsilon_{ij}$$

- **Effet Fixe (β_1)** : Cet effet représente l'impact moyen de l'activité physique sur le taux de cholestérol. Par exemple, si β_1 est négatif, cela signifie que, en moyenne, l'augmentation de l'activité physique est associée à une diminution du taux de cholestérol.
- **Effet Aléatoire (u_j)** : Cet effet reflète la variation de l'effet de l'activité physique entre les centres médicaux. Par exemple, certains centres peuvent avoir un effet plus important (ou plus faible) de l'activité physique sur la réduction du cholestérol en raison de conditions spécifiques.

Ainsi, le terme $(\beta_1 + u_j)$ représente la pente spécifique à chaque centre médical pour la relation entre l'activité physique et le taux de cholestérol. Cela permet de modéliser de manière plus réaliste la variabilité dans l'efficacité de l'activité physique à travers différents contextes.

Exemple d'Application

Pour illustrer l'effet mixte sur la covariable, imaginons qu'il y ait trois centres médicaux avec des effets différents de l'activité physique :

- **Centre A** : Forte corrélation négative entre l'activité physique et le cholestérol.
- **Centre B** : Corrélation faible.
- **Centre C** : Corrélation modérée.

Le modèle mixte permet de capturer cette hétérogénéité, en modélisant des pentes spécifiques pour chaque centre, tout en tenant compte d'une tendance globale moyenne. Cela est utile pour fournir des recommandations personnalisées en fonction des caractéristiques locales de chaque centre médical.

4. Compléments mathématiques

Les modèles mixtes peuvent être exprimés de manière compacte à l'aide d'une formulation matricielle, ce qui facilite leur interprétation et la compréhension de leur structure statistique. Une des interprétations clés de cette écriture est que le modèle mixte peut être vu comme une généralisation du modèle linéaire classique, où la matrice de covariance du bruit n'est plus diagonale. Dans ce document, nous allons détailler l'écriture matricielle des modèles mixtes et discuter de leur interprétation en termes de covariance.

Écriture Matricielle du Modèle Mixte

Considérons un modèle mixte avec des effets fixes et des effets aléatoires. Le modèle peut s'écrire sous la forme suivante :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

Où :

- $\mathbf{Y}$: Vecteur des réponses observées de taille $n \times 1$.
- $\mathbf{X}$: Matrice des prédicteurs pour les effets fixes de taille $n \times p$, avec p le nombre de prédicteurs.
- $\boldsymbol{\beta}$: Vecteur des paramètres des effets fixes de taille $p \times 1$.
- …

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

Où (suite) :

- $\mathbf{Z}$: Matrice des prédicteurs pour les effets aléatoires de taille $n \times q$, avec q le nombre d'effets aléatoires.
- $\mathbf{u}$: Vecteur des effets aléatoires de taille $q \times 1$, supposé suivre une distribution normale $\mathbf{u} \sim \mathcal{N}_q(0, \mathbf{G})$, où $\mathbf{G}$ est la matrice de covariance des effets aléatoires.
- $\boldsymbol{\varepsilon}$: Vecteur des erreurs résiduelles de taille $n \times 1$, supposé suivre une distribution normale $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{R})$, où $\mathbf{R} = \sigma^2 \mathbf{I}_n$ est la matrice de covariance des erreurs, avec $\mathbf{I}_n$ la matrice identité de taille n.

Ainsi, le modèle complet peut être décrit comme une combinaison linéaire d'effets fixes et d'effets aléatoires, avec une structure de variance-covariance complexe.

Matrice de Covariance des Réponses

Pour comprendre la structure de la variance-covariance des réponses, nous pouvons examiner la matrice de covariance de Y :

$$\text{Var}(\mathbf{Y}|\mathbf{X}, \mathbf{Z}) = \mathbf{ZGZ}' + \mathbf{R}$$

Où :

- $\mathbf{ZGZ}'$: Contribution des effets aléatoires à la variance totale. Cette matrice décrit la covariance induite par les effets aléatoires entre les observations. Elle n'est généralement pas diagonale, ce qui signifie que les observations peuvent être corrélées.
- $\mathbf{R}$: Contribution des erreurs résiduelles, souvent supposée être une matrice diagonale $\sigma^2\mathbf{I}_n$, représentant la variance individuelle de chaque observation sans corrélation entre elles.

La matrice de covariance des réponses $\text{Var}(\mathbf{Y}|\mathbf{X}, \mathbf{Z})$ n'est pas diagonale dans la plupart des cas de modèles mixtes, ce qui contraste avec le modèle linéaire classique où l'on suppose une structure de covariance simple et indépendante (i.e., matrice diagonale).

Interprétation en termes de Modèle Linéaire Classique

Dans un modèle linéaire classique, la matrice de covariance des erreurs est souvent supposée être diagonale, c'est-à-dire que les erreurs sont indépendantes et identiquement distribuées (iid). Cela implique que chaque observation est indépendante des autres. En revanche, dans un modèle mixte, la présence d'effets aléatoires implique que les erreurs ne sont plus nécessairement indépendantes, et la matrice de covariance des réponses reflète cette dépendance entre les observations.

Covariance Non Diagonale

La covariance non diagonale dans le modèle mixte est due à la présence des effets aléatoires qui introduisent des corrélations entre les réponses. Par exemple, dans le cas de données hiérarchiques, les observations au sein d'un même groupe (par exemple, les élèves dans une même école) peuvent être corrélées en raison d'un effet aléatoire associé à ce groupe (par exemple, l'effet de l'école). Ainsi, la structure de covariance de la réponse sachant les prédicteurs peut être complexe, avec des blocs de corrélations correspondant aux différents groupes.

Comparaison avec le Modèle Linéaire Classique

- **Modèle Linéaire Classique** :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

avec $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$. La matrice de covariance des erreurs est diagonale, ce qui implique des observations indépendantes.

- **Modèle Mixte** :

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

avec $\text{Var}(\mathbf{Y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$. La matrice de covariance n'est plus nécessairement diagonale, ce qui permet de capturer les corrélations entre les observations dues aux effets aléatoires.

Imaginons une étude sur les résultats des élèves dans différentes écoles, où l'effet de l'école est modélisé comme un effet aléatoire. Dans la formulation matricielle, chaque école correspond à un groupe, et les élèves au sein de chaque école partagent un effet aléatoire commun (u_j). La matrice $\mathbf{Z}\mathbf{G}\mathbf{Z}'$ introduit alors des corrélations entre les élèves d'une même école, capturant ainsi la similarité des performances au sein de chaque groupe.

Cette structure de covariance permet de mieux modéliser la variabilité présente dans les données, en tenant compte des dépendances intra-groupe qui ne pourraient pas être capturées par un modèle linéaire classique avec une matrice de covariance diagonale.

Méthodes d'Estimation des Paramètres

Pour estimer les paramètres, plusieurs méthodes peuvent être utilisées

Méthode de la Vraisemblance Maximale (ML)

La méthode de la vraisemblance maximale consiste à estimer les paramètres en maximisant la fonction de vraisemblance des observations. Cette méthode tient compte à la fois des effets fixes et aléatoires pour obtenir une estimation des paramètres du modèle.

Méthode de la Vraisemblance Restreinte (REML)

La méthode REML (Restricted Maximum Likelihood) est une variante de la vraisemblance maximale qui est souvent préférée pour les modèles mixtes. Contrairement à la méthode ML, REML ne se concentre que sur la partie de la vraisemblance qui ne dépend pas des effets fixes, ce qui permet de réduire le biais des estimations des variances des effets aléatoires.